

Variable selection and structure identification for varying coefficient Cox models

Toshio Honda*

*Graduate School of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601,
Japan*

Ryota Yabe

Department of Economics, Shinshu University, Matsumoto, Nagano 390-8621, Japan

Abstract

We consider varying coefficient Cox models with high-dimensional covariates. We apply the group Lasso method to these models and propose a variable selection procedure. Our procedure copes with variable selection and structure identification from a high dimensional varying coefficient model to a semivarying coefficient model simultaneously. We derive an oracle inequality and closely examine restrictive eigenvalue conditions, too. In this paper, we give the details for Cox models with time-varying coefficients. The theoretical results on variable selection can be easily extended to some other important models and we briefly mention those models since those models can be treated in the same way. The models considered in this paper are the most popular models among structured nonparametric regression models. The results of a small numerical study are also given.

Keywords: censored survival data, high-dimensional data, group Lasso, B-spline basis, structured nonparametric regression model, semivarying coefficient model

2010 MSC: 62G08, 62N01

*Corresponding author

Email addresses: t.honda@r.hit-u.ac.jp (Toshio Honda),
ryotayabe@shinshu-u.ac.jp (Ryota Yabe)

1. Introduction

The Cox model is one of the most popular and useful models to analyze censored survival data. Since the Cox model was proposed in Cox[8], many authors have studied a lot of extensions or variants of the original Cox model to deal with complicated situations or carry out more flexible statistical analysis. In this paper, we consider varying coefficient models and additive models with high-dimensional covariates. These models with moderate numbers of covariates are investigated in many papers, for example, Huang et al.[13], Cai and Sun[7], and Cai et al.[6].

We apply the group Lasso (for example, see Lounici et al.[17]) to varying coefficient models with high-dimensional covariates to carry out variable selection and structure identification simultaneously. Although we focus on time-varying coefficient models here, our method can be applied to variable selection for another type of varying coefficient models and additive models and we briefly mention how to apply and how to derive the theoretical results.

Suppose that we observe censored survival times T_i and high-dimensional covariates $\mathbf{X}_i(t) = (X_{i1}(t), \dots, X_{ip}(t))^T$. More specifically, we have n i.i.d. observations of

$$T_i = \min\{T_{0i}, C_i\}, \quad \delta_i = I\{T_{0i} \leq C_i\}, \quad (1)$$

and p -dimensional covariate $\mathbf{X}_i(t)$ on the time interval $[0, \tau]$, where T_{0i} is an uncensored survival time and C_i is a censoring time satisfying the condition of the independent censoring mechanism as in section 6.2 of Kalbfleisch and Prentice[15]. Hereafter we set $\tau = 1$ for simplicity of presentation. Note that p can be very large compared to n in this paper, for example, $p = O(n^{c_p})$ for a very large positive constant c_p or $p = O(\exp(n^{c_p}))$ for a sufficiently small positive constant c_p . We assume that the standard setup for the Cox model holds as in chapter 5 of [15] and that T_i or $N_i(t) = I\{t \geq T_i\}$ has the following compensator $\Lambda_i(t)$ with respect to a suitable filtration $\{\mathcal{F}_t\}$:

$$d\Lambda_i(t) = Y_i(t) \exp\{\mathbf{X}_i(t)^T \mathbf{g}(t)\} \lambda_0(t) dt, \quad (2)$$

where $Y_i(t) = I\{t \leq T_i\}$, $\mathbf{g}(t) = (g_1(t), \dots, g_p(t))^T$ is a vector of unknown functions on $[0, 1]$, \mathbf{a}^T denotes the transpose of \mathbf{a} , and $\lambda_0(t)$ is a baseline hazard function. As in chapter 5 of [15], $\mathbf{X}_i(t)$ is predictable and

$$M_i(t) = N_i(t) - \Lambda_i(t) \quad (3)$$

is a martingale with respect to $\{\mathcal{F}_t\}$. In the original Cox model, $\mathbf{g}(t)$ is a vector of constants and we estimate this constant coefficient vector by maximizing the partial likelihood.

In this paper, we are interested in estimating $\mathbf{g}(t)$ in (2). Recently we have many cases where there are (ultra) high-dimensional covariates due to drastic development of data collecting technology. In such high-dimensional data, usually only a small part of covariates are relevant. However, we cannot directly apply standard or traditional estimating procedures to such high-dimensional data. Thus now a lot of methods for variable selection are available, for example, SCAD and Lasso procedures. See Bühlmann and van de Geer[5] and Hastie et al.[10] for excellent reviews of these procedures for variable selection. See also Bickel et al.[2] and Zou[29] for the Lasso and the adaptive Lasso, respectively.

As for high dimensional Cox models with constant coefficient, Bradic et al.[3] studied the SCAD method and Huang et al.[12] considered the Lasso procedure. The authors of [12] developed new ingenious techniques to derive oracle inequalities. We will fully use their techniques to derive our theoretical results such as an oracle inequality. In addition, Zhang and Luo[24] proposed an adaptive Lasso estimator for the Cox model. Some variable screening procedures have also been proposed in Zhao and Li[28] and Yang et al.[22], to name just a few.

In this paper, we propose a group Lasso procedure to select relevant covariates and identify the covariates with constant coefficients among the relevant covariates, namely the true semivarying coefficient model from the original varying coefficient model. We can achieve this goal by the proposed group Lasso with a suitable threshold value or a two-stage procedure consisting of the proposed one and an adaptive Lasso procedure as in Yan and Huang[21] and Honda and Härdle[11]. In [21], the authors proposed an adaptive Lasso procedure for structure identification with no theoretical result. Our procedure can be applied to the varying coefficient model with an index variable $Z_i(t)$:

$$d\Lambda_i(t) = Y_i(t) \exp\{g_0(Z_i(t)) + \mathbf{X}_i(t)^T \mathbf{g}(Z_i(t))\} \lambda_0(t) dt \quad (4)$$

and the additive model:

$$d\Lambda_i(t) = Y_i(t) \exp\left\{\sum_{j=1}^p g_j(X_{ij}(t))\right\} \lambda_0(t) dt. \quad (5)$$

We mention these model later in section 4.

Some authors considered the same problem by using SCAD. For example, see Lian et al.[16] and Zhang et al.[25]. They proved the existence of local optimizer satisfying the same convergence rate as ours. In contrast, we prove the existence of the global solution with desirable properties. In Bradic and Song[4], the authors applied penalties similar to ours to additive models and obtained theoretical results in another complicated manner. We have derived a better convergence rate for our procedures to varying coefficient models by exploiting the martingale structure very carefully under much simpler assumptions given in section 2. See Remark 1 in section 3 about the convergence rate. We also carefully examined the RE (restrictive eigenvalue) conditions. While the other authors considered the L_2 norm of the estimated second derivatives for additive models, we adopt the orthogonal decomposition approach. We give some details on why we have adopted the orthogonal decomposition approach in Appendix D.

This paper is organized as follows. In section 2, we describe our group Lasso procedure for time-varying coefficient models. Then we present our theoretical results in section 3. We mention the two other models in section 4. The results of a small simulation study are given in section 5. The proofs of our theoretical results are postponed to section 6 and section 7 concludes this paper. We collected useful properties of our basis functions and the proofs of technical lemmas in Appendices A-E.

We define some notation and symbols here. In this paper, C, C_1, C_2, \dots are positive generic constants and their values change from line to line. For a vector \mathbf{a} , $|\mathbf{a}|$, $|\mathbf{a}|_1$, and $|\mathbf{a}|_\infty$ mean the L_2 norm, the L_1 norm, and the sup norm, respectively. For a function g on $[0, 1]$, $\|g\|$, $\|g\|_1$, and $\|g\|_\infty$ stand for the L_2 norm, the L_1 norm, and the sup norm, respectively. For a symmetric matrix A , we denote the minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Besides, $\text{sign}(a)$ is the sign of a real number a and $a_n \sim b_n$ means there are positive constants C_1 and C_2 such that $C_1 < a_n/b_n < C_2$. We write $\overline{\mathcal{S}}$ for the complement of a set \mathcal{S} .

2. Group Lasso procedure

First we decompose $g_j(t)$, $j = 1, \dots, p$, into the constant part and the non-constant part:

$$g_j(t) = g_{cj} + g_{nj}(t), \quad (6)$$

where $\int_0^1 g_{nj}(t)dt = 0$. When $g_j(t) \not\equiv 0$, $g_j(t)$ is a non-zero constant or a non-constant function. We denote the index sets of relevant covariates by

$$\mathcal{S}_c = \{j \mid g_{cj} \neq 0\} \quad \text{and} \quad \mathcal{S}_n = \{j \mid g_{nj}(t) \not\equiv 0\} \quad (7)$$

and set

$$s_c = \#\mathcal{S}_c, \quad s_n = \#\mathcal{S}_n, \quad \text{and} \quad s_o = s_c + s_n,$$

where $\#A$ is the number of the elements of a set A . We implicitly assume that s_o is bounded or much smaller than n . Besides, we assume

$$\mathcal{S}_n \subset \mathcal{S}_c. \quad (8)$$

We may incidentally have $g_{cj} = 0$ for $j \in \mathcal{S}_n$. However, this will rarely happen and g_{cj} should be free if $g_{nj}(t) \not\equiv 0$.

Next we introduce our spline basis $\overline{\mathbf{B}}(t)$ to approximate $g_j(t)$, $j = 1, \dots, p$. We construct $\overline{\mathbf{B}}(t)$ from the L -dimensional equispaced B-spline basis $\mathbf{B}_0(t) = (b_{01}(t), \dots, b_{0L}(t))^T$ on $[0, 1]$ and the basis has the following properties :

$$\overline{\mathbf{B}}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_L(t) \end{pmatrix} = \begin{pmatrix} 1/\sqrt{L} \\ \mathbf{B}(t) \end{pmatrix} = A_0 \mathbf{B}_0(t) \quad \text{and} \quad \int_0^1 \overline{\mathbf{B}}(t) \overline{\mathbf{B}}^T(t) dt = L^{-1} I, \quad (9)$$

where

$$A_0 = \begin{pmatrix} \mathbf{a}_{01}^T \\ \mathbf{a}_{02}^T \\ \vdots \\ \mathbf{a}_{0L}^T \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T/\sqrt{L} \\ A_{-1} \end{pmatrix} \quad (\text{say})$$

and $\mathbf{1} = (1, \dots, 1)^T$. Note that for $j = 1, \dots, L$,

$$b_j(t) = \mathbf{a}_{0j}^T \mathbf{B}_0(t)$$

and that $1/\sqrt{L}$ and $\mathbf{B}(T) = (b_2(t), \dots, b_L(t))^T$ in (9) are designed for g_{cj} and $g_{nj}(t)$, respectively. Recall that $\mathbf{1}^T \mathbf{B}_0(t) \equiv 1$ and see Schumaker[18] for the definition of B-spline bases. We have collected how to construct $\overline{\mathbf{B}}(t)$ and A_0 and some useful properties of $\overline{\mathbf{B}}(t)$ and A_0 in Appendix A. We can use another basis which has desirable properties such as (A.1), (A.3), and (A.4) in Appendix A.

We impose some technical assumptions on $\mathbf{g}(t)$.

Assumption G : $g_j(t)$, $j = 1, \dots, p$, are twice continuously differentiable and there is a positive constant C_g such that

$$\sum_{j=1}^p \|g_j\|_\infty \leq C_g, \quad \sum_{j=1}^p \|g'_j\|_\infty \leq C_g, \quad \text{and} \quad \sum_{j=1}^p \|g''_j\|_\infty \leq C_g.$$

Besides we have

$$\min_{j \in \mathcal{S}_c} |g_{cj}| L^2 \rightarrow \infty \quad \text{and} \quad \min_{j \in \mathcal{S}_n} \|g_{nj}\| L^2 \rightarrow \infty.$$

Hereafter we take $L = c_L n^{1/5}$ ($c_L > 0$) for simplicity of presentation and the order of the B-spline basis should be larger than or equal to 2. The latter of Assumption G means relevant coefficient functions are larger than the approximation error. As for the identifiability of $\mathbf{g}(t)$, we need an assumption such as $\lambda_{\min}(\mathbb{E}\{\bar{\Sigma}\}) > C_1/L$ for a positive constant C_1 , where $\mathbb{E}\{\bar{\Sigma}\}$ is defined in Proposition 3.

When Assumption G holds, there are $\boldsymbol{\gamma}_j^* = (\gamma_{1j}^*, \gamma_{-1j}^{*T})^T \in R^L$, $j = 1, \dots, p$, such that for a positive constant C_{approx} depending on C_g ,

$$\sum_{j=1}^p \|g_j - \bar{\mathbf{B}}(t)^T \boldsymbol{\gamma}_j^*\|_\infty \leq C_{approx} L^{-2}. \quad (10)$$

When $j \in \mathcal{S}_c$, we can take $\gamma_{1j}^* = \sqrt{L} g_{cj}$ and $\boldsymbol{\gamma}_{-1j}^* \in R^{L-1}$ depends on $g_{jn}(t)$. If $j \in \bar{\mathcal{S}}_n$, we take $\boldsymbol{\gamma}_{-1j}^* = 0$. When $j \in \bar{\mathcal{S}}_c$, we set $\boldsymbol{\gamma}_j^* = 0$. See Appendix A for more details on these $\boldsymbol{\gamma}_j^* = (\gamma_{1j}^*, \boldsymbol{\gamma}_{-1j}^{*T})^T$.

We state assumptions on our Cox model before we describe the log partial likelihood for new covariates

$$\mathbf{W}_i(t) = \mathbf{X}_i(t) \otimes \bar{\mathbf{B}}(t), \quad (11)$$

where \otimes means the Kronecker product.

Assumption M : $|X_{1j}(t)| \leq C_X$ uniformly in j and t for a positive constant C_X . We also have $\mathbb{E}\{Y_1(1)\} \geq C_Y$ for a positive constant C_Y . Besides, the baseline hazard function is bounded from above and satisfies $\lambda_0(t) \geq C_\lambda$ on $[0, 1]$ for a positive constant C_λ .

The first one is used to evaluate the inside of the exponential function and the other ones are standard in the literature.

We denote the log partial likelihood by $L_p(\boldsymbol{\gamma})$:

$$L_p(\boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^n \int_0^1 \boldsymbol{\gamma}^T \mathbf{W}_i(t) dN_i(t) - \int_0^1 \log \left[\sum_{i=1}^n Y_i(t) \exp\{\boldsymbol{\gamma}^T \mathbf{W}_i(t)\} \right] d\overline{N}(t), \quad (12)$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_p^T)^T \in R^{pL}$ and $\overline{N}(t) = n^{-1} \sum_{i=1}^n N_i(t)$. We also use the same sample mean notation for $M_i(t)$ and $Y_i(t)$.

Set

$$l_p(\boldsymbol{\gamma}) = -L_p(\boldsymbol{\gamma}) \quad (13)$$

for notational convenience. Then we should minimize this $l_p(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. However, when pL is larger than n , we cannot carry out this minimization properly and we add some penalty as in the literature on high-dimensional data. We define two convex penalties here :

$$P_1(\boldsymbol{\gamma}) = \sum_{j=1}^p (|\gamma_{1j}| + |\gamma_{-1j}|) \quad (14)$$

and

$$P_h(\boldsymbol{\gamma}) = \sum_{j=1}^p (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} + \sum_{j=1}^p |\gamma_{-1j}| \quad (15)$$

for some $q > 1$.

This $P_1(\boldsymbol{\gamma})$ plays the role of the L_1 norm for $\boldsymbol{\gamma} \in R^{pL}$ and is a very important technical tool in this paper. Besides, we define a kind of sup norm $P_\infty(\boldsymbol{\gamma})$ by

$$P_\infty(\boldsymbol{\gamma}) = \max_{1 \leq j \leq p} |\gamma_{1j}| \vee |\gamma_{-1j}|, \quad (16)$$

where $a \vee b = \max\{a, b\}$. This is also an important tool.

We defined the penalty in (15) by taking the assumption in (8) into consideration and following Zhao et al.[27] and Zhao and Leng[26]. Thus our group Lasso objective functions are

$$Q_h(\boldsymbol{\gamma}; \lambda) = l_p(\boldsymbol{\gamma}) + \lambda P_h(\boldsymbol{\gamma}) \quad \text{and} \quad Q_1(\boldsymbol{\gamma}; \lambda) = l_p(\boldsymbol{\gamma}) + \lambda P_1(\boldsymbol{\gamma}). \quad (17)$$

Our group Lasso estimate is given by

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in R^{pL}}{\operatorname{argmin}} Q_h(\boldsymbol{\gamma}; \lambda) \quad \text{or} \quad \hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in R^{pL}}{\operatorname{argmin}} Q_1(\boldsymbol{\gamma}; \lambda).$$

If we are interested in only variable selection, we should minimize

$$Q(\boldsymbol{\gamma}; \lambda) = l_p(\boldsymbol{\gamma}) + \lambda \sum_{j=1}^p |\gamma_j|. \quad (18)$$

The KKT condition implies that for $a = h$ or 1,

$$\frac{\partial l_p}{\partial \gamma_j}(\hat{\boldsymbol{\gamma}}) = -\lambda \nabla_j P_a(\hat{\boldsymbol{\gamma}}), \quad j = 1, \dots, p, \quad (19)$$

where $\nabla_j P_a(\boldsymbol{\gamma})$ is the subgradient of $P_a(\boldsymbol{\gamma})$ with respect to γ_j . See chapter 5 of [10] about convex optimality conditions. We give explicit expressions of these subgradients in Appendix B for reference. Consequently from (9), our estimates of g_{cj} and g_{nj} are

$$\hat{g}_{cj} = \hat{\gamma}_{1j}/\sqrt{L} \quad \text{and} \quad \hat{g}_{nj}(t) = \mathbf{B}^T(t)\hat{\boldsymbol{\gamma}}_{-1j}. \quad (20)$$

If we choose a threshold value t_λ based on our theoretical results in section 3 and define $\hat{\mathcal{S}}_c$ and $\hat{\mathcal{S}}_n$ by

$$\hat{\mathcal{S}}_c = \{j \mid |\hat{g}_{cj}| > t_\lambda\} \quad \text{and} \quad \hat{\mathcal{S}}_n = \{j \mid \|\hat{g}_{nj}\| > t_\lambda\}, \quad (21)$$

they are consistent estimators of \mathcal{S}_c and \mathcal{S}_n , respectively. Or we can apply an adaptive Lasso procedure to estimate the true semivarying coefficient model.

We state our theoretical results only for $Q_h(\boldsymbol{\gamma}; \lambda)$ in section 3 since we can deal with $Q_1(\boldsymbol{\gamma}; \lambda)$ and $Q(\boldsymbol{\gamma}; \lambda)$ in the same way. In terms of numerical optimization, $Q_1(\boldsymbol{\gamma}; \lambda)$ seems to be more tractable and we focused on $Q_1(\boldsymbol{\gamma}; \lambda)$ in our numerical study. When the group Lasso based on $Q_1(\boldsymbol{\gamma}; \lambda)$ concludes that $\|g_{nj}\| > 0$ and $|g_{cj}| = 0$, we should take (8) into consideration and modify this conclusion to the one that both of them are relevant for this j .

3. Oracle inequality

An oracle inequality for $\hat{\boldsymbol{\gamma}}$ from $Q_h(\boldsymbol{\gamma}; \lambda)$ is given in Theorem 1. First we define some notation. We borrow some notation from [12] and proceed as in [12]. Some other notation is standard in the literature of the Cox model and the Lasso.

Let $\boldsymbol{\gamma}_S$ consist of $\{\gamma_{1j}\}_{j \in \mathcal{S}_c}$ and $\{\gamma_{-1j}\}_{j \in \mathcal{S}_n}$. On the other hand, $\boldsymbol{\gamma}_{\bar{S}}$ consists of $\{\gamma_{1j}\}_{j \in \bar{\mathcal{S}}_c}$ and $\{\gamma_{-1j}\}_{j \in \bar{\mathcal{S}}_n}$.

We need some notation to give explicit expressions of the derivatives of $l_p(\gamma)$.

$$S^{(k)}(t, \gamma) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{W}_i^{\otimes k}(t) \exp\{\mathbf{W}_i^T(t) \gamma\},$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a} \mathbf{a}^T$. In addition,

$$\widetilde{\mathbf{W}}_n(t, \gamma) = \frac{S^{(1)}(t, \gamma)}{S^{(0)}(t, \gamma)} \quad \text{and} \quad V_n(t, \gamma) = \frac{S^{(2)}(t, \gamma)}{S^{(0)}(t, \gamma)} - (\widetilde{\mathbf{W}}_n(t, \gamma))^{\otimes 2}. \quad (22)$$

Hence we have the following expressions of the derivatives of $l_p(\gamma)$, which are denoted by $\dot{l}_p(\gamma)$ and $\ddot{l}_p(\gamma)$:

$$\frac{\partial l_p}{\partial \gamma}(\gamma) = -\frac{1}{n} \sum_{i=1}^n \int_0^1 \{\mathbf{W}_i(t) - \widetilde{\mathbf{W}}_n(t, \gamma)\} dN_i(t) = \dot{l}_p(\gamma) \quad (\text{say}) \quad (23)$$

and

$$\frac{\partial^2 l_p}{\partial \gamma \partial \gamma^T}(\gamma) = \int_0^1 V_n(t, \gamma) d\overline{N}(t) = \ddot{l}_p(\gamma) \quad (\text{say}). \quad (24)$$

In Proposition 1, we prove that $\widehat{\gamma}$ is in a restricted parameter space. We define some more notation to state Proposition 1. Set

$$D_l = P_\infty(\dot{l}_p(\gamma^*)) \quad \text{and} \quad \widehat{\boldsymbol{\theta}} = \widehat{\gamma} - \gamma^*. \quad (25)$$

We evaluate D_l later in Proposition 2. We define $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_{\overline{S}}$ in the same way as γ_S and $\gamma_{\overline{S}}$. Recall that $\gamma^* = (\gamma_1^{*T}, \dots, \gamma_p^{*T})^T$ is given in (10). This proposition follows from only (19).

Proposition 1. *If $\lambda > D_l$, we have*

$$(\widehat{\gamma} - \gamma^*)^T \{\dot{l}_p(\widehat{\gamma}) - \dot{l}_p(\gamma^*)\} \leq (2\lambda + D_l) P_1(\widehat{\boldsymbol{\theta}}_S) - (\lambda - D_l) P_1(\widehat{\boldsymbol{\theta}}_{\overline{S}})$$

and

$$(\lambda - D_l) P_1(\widehat{\boldsymbol{\theta}}_{\overline{S}}) \leq (2\lambda + D_l) P_1(\widehat{\boldsymbol{\theta}}_S).$$

Therefore if $D_l \leq \xi \lambda$ ($\xi < 1$), we have

$$P_1(\widehat{\boldsymbol{\theta}}_{\overline{S}}) \leq \frac{2 + \xi}{1 - \xi} P_1(\widehat{\boldsymbol{\theta}}_S).$$

We define a restricted parameter space $\Theta(\zeta)$ by

$$\Theta(\zeta) = \{\boldsymbol{\theta} \in R^{pL} \mid P_1(\boldsymbol{\theta}_{\mathcal{S}}) \leq \zeta P_1(\boldsymbol{\theta}_{\mathcal{S}})\}.$$

For $\boldsymbol{\theta} \in \Theta(\zeta)$, we have

$$P_1(\boldsymbol{\theta}) \leq (1 + \zeta)P_1(\boldsymbol{\theta}_{\mathcal{S}}) \quad \text{and} \quad P_1(\boldsymbol{\theta}_{\mathcal{S}}) \leq s_0^{1/2}|\boldsymbol{\theta}_{\mathcal{S}}| \leq s_0^{1/2}|\boldsymbol{\theta}|. \quad (26)$$

Recall that s_0 is defined just after (7).

To state the compatibility and restrictive eigenvalue conditions, we define $\kappa(\zeta, \Sigma)$ and $RE(\zeta, \Sigma)$ for an n.n.d. (non-negative definite) matrix Σ with some modifications adapted to our setup.

$$\kappa(\zeta, \Sigma) = \inf_{\boldsymbol{\theta} \in \Theta(\zeta), \boldsymbol{\theta} \neq 0} \frac{s_0^{1/2}(\boldsymbol{\theta}^T \Sigma \boldsymbol{\theta})^{1/2}}{P_1(\boldsymbol{\theta}_{\mathcal{S}})} \quad \text{and} \quad RE(\zeta, \Sigma) = \inf_{\boldsymbol{\theta} \in \Theta(\zeta), \boldsymbol{\theta} \neq 0} \frac{(\boldsymbol{\theta}^T \Sigma \boldsymbol{\theta})^{1/2}}{|\boldsymbol{\theta}|}.$$

The latter is more commonly used in the literature of the Lasso. It is known that

$$\kappa^2(\zeta, \Sigma) \geq RE^2(\zeta, \Sigma) \geq \lambda_{\min}(\Sigma)$$

and that if $\Sigma_1 - \Sigma_2$ is n.n.d., we also have

$$\kappa(\zeta, \Sigma_1) \geq \kappa(\zeta, \Sigma_2) \quad \text{and} \quad RE(\zeta, \Sigma_1) \geq RE(\zeta, \Sigma_2).$$

Some more notation is necessary for Theorem 1. Set

$$C_W = 2C_X \{\lambda_{\max}(A_0 A_0^T)\}^{1/2}, \quad RE^* = RE\left(\frac{2+\xi}{1-\xi}, \ddot{l}_p(\boldsymbol{\gamma}^*)\right), \quad (27)$$

$$\kappa^* = \kappa\left(\frac{2+\xi}{1-\xi}, \ddot{l}_p(\boldsymbol{\gamma}^*)\right), \quad \text{and} \quad \tau^* = \frac{9s_0 \lambda C_W}{4(1-\xi)(\kappa^*)^2} \quad \text{for } \xi \in (0, 1). \quad (28)$$

Note that C_W is bounded from above. We closely look at RE^* and κ^* in Proposition 3. Let η^* be the smaller solution of

$$\eta \exp(-\eta) = \tau^*$$

as in [12]. Note that τ^* should tend to 0 as in Remark 1.

Recall that we are considering $Q_h(\boldsymbol{\gamma}; \lambda)$ now since we can deal with $Q_1(\boldsymbol{\gamma}; \lambda)$ in (17) and $Q(\boldsymbol{\gamma}; \lambda)$ in (18) in almost the same way and drive the same results with just conformable changes.

Theorem 1. Assume that Assumptions G and M hold. Then if $D_l \leq \xi\lambda$ for some $\xi \in (0, 1)$, we have

$$P_1(\widehat{\gamma} - \gamma^*) \leq \eta^*/C_W.$$

Then we also have

$$\begin{aligned} \max_{1 \leq j \leq p} |\widehat{g}_{cj} - g_{cj}| &\leq C_c \left(\frac{\eta^*}{L^{1/2}} + L^{-2} \right), \quad \max_{1 \leq j \leq p} \|\widehat{g}_{nj} - g_{nj}\| \leq C_{n1} \left(\frac{\eta^*}{L^{1/2}} + L^{-2} \right), \\ \max_{1 \leq j \leq p} \|\widehat{g}_{nj} - g_{nj}\|_\infty &\leq C_{n2} \left(\frac{\eta^*}{L^{1/2}} + L^{-2} \right), \end{aligned}$$

where C_c , C_{n1} , and C_{n2} depend on C_W , C_g , and the properties of the B-spline basis on $[0, 1]$ and they are bounded.

Some remarks are in order.

Remark 1. When $p = O(n^{c_p})$ for some c_p , we have $D_l = O_p((n^{-1} \log n)^{1/2})$ and should take $\lambda = C(n^{-1} \log n)^{1/2}$ for some sufficiently large C . As in shown in Proposition 3, we usually have $(\kappa^*)^2 \sim L^{-1}$ with probability tending to 1 in suitable setups. Then $\tau^* \sim L(n^{-1} \log n)^{1/2}$ and $\eta^*/\tau^* \rightarrow 1$. This leads to the convergence rate of $O(n^{-2/5}(\log n)^{1/2})$ for \widehat{g}_{cj} and \widehat{g}_{nj} and improves that of [4], which is $O(n^{-7/20}(\log n)^{1/2})$ for their additive model in a similar setup. Our rate is optimal except for $(\log n)^{1/2}$. Our results can deal with ultra high-dimensional cases if $p = \exp(c_p n)$ and c_p is sufficiently small. See Propositions 2 and 3.

Remark 2. Suppose that

$$\max_{j \in \mathcal{S}_c} |g_{cj}| / (n^{-2/5}(\log n)^{1/2}) \rightarrow \infty \quad \text{and} \quad \max_{j \in \mathcal{S}_n} \|g_{nj}\| / (n^{-2/5}(\log n)^{1/2}) \rightarrow \infty.$$

Then if we take t_λ satisfying $t_\lambda/\lambda \rightarrow \infty$ sufficiently slowly for λ in Remark 1, $\widehat{\mathcal{S}}_c$ and $\widehat{\mathcal{S}}_n$ in (21) are consistent estimators of \mathcal{S}_c and \mathcal{S}_n , respectively.

Next we evaluate D_l in Proposition 2, which is called the deviation condition. From Assumption M and application of Bernstein's inequality (for example, see [20]), we have with probability larger than $1 - P_Y$,

$$\frac{1}{n} \sum_{i=1}^n Y_i(1) = \bar{Y}(1) > C_Y, \tag{29}$$

where

$$P_Y = \exp \left\{ - \frac{C_Y^2 n}{2(1 + 2C_Y/3)} \right\}.$$

Since

$$\dot{l}_p(\gamma^*) = -\frac{1}{n} \sum_{i=1}^n \int_0^1 \{ \mathbf{W}_i(t) - \widetilde{\mathbf{W}}_n(t, \gamma^*) \} dN_i(t), \quad (30)$$

we evaluate \dot{l}_{op} in (31) and $\dot{l}_{op} - \dot{l}_p(\gamma^*)$ in (32).

$$\begin{aligned} \dot{l}_{op} &= -\frac{1}{n} \sum_{i=1}^n \int_0^1 \left\{ \mathbf{W}_i(t) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)} \right\} dN_i(t) \\ &= -\frac{1}{n} \sum_{i=1}^n \int_0^1 \left\{ \mathbf{W}_i(t) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)} \right\} dM_i(t), \end{aligned} \quad (31)$$

where

$$\begin{aligned} S_0^{(k)}(t) &= \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{W}_i^{\otimes k}(t) \exp\{\mathbf{g}^T(t) \mathbf{X}_i(t)\}, \quad k = 0, 1, 2. \\ \dot{l}_{op} - \dot{l}_p(\gamma^*) &= \int_0^1 \left\{ \widetilde{\mathbf{W}}_n(t, \gamma^*) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)} \right\} d\bar{N}(t). \end{aligned} \quad (32)$$

By combining evaluations of (31) and (32), we obtain Proposition 2. The proof is postponed to section 6. Recall that $\widetilde{\mathbf{W}}_n(t, \gamma^*)$ is defined in (22).

Proposition 2. *Assume that Assumptions G and M hold. Then we have*

$$P_\infty(\dot{l}_p(\gamma^*)) \leq \frac{a_1}{L^{5/2}} + \frac{x(\log n)^{1/2}}{\sqrt{n}}$$

with probability larger than

$$1 - P_Y - La_2 \exp\{-a_3 n L^{-1}\} - 2pL \exp \left\{ - \frac{a_4 x^2 \log n}{1 + x(n^{-1} L \log n)^{1/2}} \right\},$$

where a_j , $j = 1, \dots, 4$, are positive constants depending only on the assumptions and they are independent of n .

Finally we deal with κ^* and RE^* . In Proposition 3, we give their lower bounds. They are called the compatibility condition and the restricted eigenvalue condition, respectively.

Proposition 3. *Assume that Assumptions G and M hold. Then with probability larger than $1 - P_Y - P_A - P_B - P_C$, we have*

$$\begin{aligned} \kappa^2(\zeta, \ddot{l}_p(\gamma)) &\geq \exp(-C_X C_g)(1 + O(L^{-2}))\kappa^2(\zeta, E\{\bar{\Sigma}\}) \\ &\quad - s_0(1 + \zeta)^2 L \left\{ \frac{c_1}{L^3} + \frac{x(\log n)^{1/2}}{\sqrt{nL}} \right\} \end{aligned}$$

and

$$\begin{aligned} RE^2(\zeta, \ddot{l}_p(\gamma)) &\geq \exp(-C_X C_g)(1 + O(L^{-2}))RE^2(\zeta, E\{\bar{\Sigma}\}) \\ &\quad - s_0(1 + \zeta)^2 L \left\{ \frac{c_2}{L^3} + \frac{x(\log n)^{1/2}}{\sqrt{nL}} \right\} \end{aligned}$$

where

$$\begin{aligned} \bar{\Sigma} &= \int_0^1 \bar{G}_Y(t) \lambda_0(t) dt, \quad \bar{G}_Y(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \{\mathbf{W}_i(t) - \boldsymbol{\mu}_Y(t)\}^{\otimes 2}, \\ \boldsymbol{\mu}_Y(t) &= \frac{E\{Y_1(t) \mathbf{W}_1(t)\}}{E\{Y_1(t)\}}, \quad P_A = 2(pL)^2 \exp \left\{ - \frac{c_3 x^2 \log n}{1 + x(\log n)^{1/2} (n^{-1} L)^{1/2}} \right\}, \\ P_B &= 5(pL)^2 \exp \left\{ - \frac{c_4 x (n \log n)^{1/2}}{1 + x^{1/2} (n^{-1} \log n)^{1/4}} \right\}, \\ P_C &= 2(pL)^2 \exp \left\{ - \frac{c_5 x^2 \log n}{1 + x(\log n)^{1/2} n^{-1}} \right\}. \end{aligned}$$

Note that c_j , $j = 1, \dots, 5$, are positive constants depending only on the assumptions and they are independent of n .

In the literature, it is often assumed that there is a positive constant C_1 such that $\lambda_{\min}(E\{\bar{\Sigma}\}) \geq C_1/L$ due to (A.1) and (A.2) in Appendix A. Then for some positive constants C_2 and C_3 , we have

$$\kappa^2(\zeta, E\{\bar{\Sigma}\}) \geq \frac{C_2}{L} + o_p(L^{-1}) \quad \text{and} \quad RE^2(\zeta, E\{\bar{\Sigma}\}) \geq \frac{C_3}{L} + o_p(L^{-1})$$

if s_0 is bounded and $p = O(n^{c_p})$.

4. Other models

4.1. Varying coefficient models with index variables

When we observe $(Z_i(t), \mathbf{X}_i(t))$ and $Z_i(t)$ is an influential variable treated as the index variable, the following model for the compensator is among candidates of our models for statistical analysis.

$$d\Lambda_i(t) = Y_i(t) \exp\{g_0(Z_i(t)) + \mathbf{X}_i(t)^T \mathbf{g}(Z_i(t))\} \lambda_0(t) dt, \quad (33)$$

where $Z_i(t) \in [0, 1]$, $\int_0^1 g_0(z) dz = 0$, and $g_j(z) = g_{cj} + g_{nj}(z)$, $j = 1, \dots, p$, as in section 2. Then we can proceed in almost the same way with

$$\begin{aligned} \mathbf{W}_i(t) &= (\mathbf{B}^T(Z_i(t)), \mathbf{X}_i(t)^T \otimes \overline{\mathbf{B}}^T(Z_i(t)))^T, \\ \boldsymbol{\gamma} &= (\boldsymbol{\gamma}_{-10}^T, \gamma_{11}, \boldsymbol{\gamma}_{-11}^T, \dots, \gamma_{1p}, \boldsymbol{\gamma}_{-1p}^T)^T, \\ P_1(\boldsymbol{\gamma}) &= \sum_{j=0}^p |\gamma_{1j}| + \sum_{j=1}^p |\boldsymbol{\gamma}_{-1j}|, \\ P_h(\boldsymbol{\gamma}) &= \sum_{j=1}^p (|\gamma_{1j}|^q + |\boldsymbol{\gamma}_{-1j}|^q)^{1/q} + \sum_{j=0}^p |\boldsymbol{\gamma}_{-1j}|, \\ P_\infty(\boldsymbol{\gamma}) &= \{\max_{1 \leq j \leq p} |\gamma_{1j}| \vee |\boldsymbol{\gamma}_{-1j}|\} \vee |\boldsymbol{\gamma}_{-10}|, \\ Q_1(\boldsymbol{\gamma}; \lambda) &= l_p(\boldsymbol{\gamma}) + \lambda P_1(\boldsymbol{\gamma}), \quad \text{and} \quad Q_h(\boldsymbol{\gamma}; \lambda) = l_p(\boldsymbol{\gamma}) + \lambda P_h(\boldsymbol{\gamma}). \end{aligned}$$

We can carry out simultaneous variable selection and structure identification of this model as for time-varying coefficient models and we are able to prove the same results in almost the same way. Almost no change is necessary to the proofs of Proposition 1 and Theorem 1. When we consider Propositions 2 and 3, we should be a little careful in evaluating predictable variation processes and so on. Then we have to deal with terms like

$$n^{-1} \sum_{i=1}^n |b_{0j}(Z_i(t))|, \quad n^{-1} \sum_{i=1}^n |b_j(Z_i(t))|, \quad \text{and} \quad n^{-1} \sum_{i=1}^n |b_j(Z_i(t)) b_k(Z_i(t))|$$

as compared to

$$|b_{0j}(t)|, \quad |b_j(t)|, \quad \text{and} \quad |b_j(t) b_k(t)|$$

for time-varying coefficient models. Note that we can use exponential inequalities for generalized U-statistics as given in Gine et al.[9] instead of Lemma 4.2 in [12] in the proof of Proposition 3. We give more details in Appendix E.

4.2. Additive models

When we have no specific index variable, the following additive model may be suitable.

$$d\Lambda_i(t) = Y_i(t) \exp \left\{ \sum_{j=1}^p g_j(X_{ij}(t)) \right\} \lambda_0(t) dt, \quad (34)$$

where $\int_0^1 g_j(x) dx = 0$ and $X_{ij}(t) \in [0, 1]$. These $g_j(x)$ can be orthogonally decomposed into the linear part and the nonlinear part as well.

We should take $b_2(X_{ij}(t)) = (12L^{-1})^{1/2}(X_{ij}(t) - 1/2)$ and use $b_2(X_{ij}(t))$ and $(b_3(X_{ij}(t)), \dots, b_L(X_{ij}(t)))^T$ for the linear part and the nonlinear part, respectively. We have no $b_1(X_{ij}(t))$ and divide γ_{-1j} into γ_{2j} and $\gamma_{-2j} = (\gamma_{3j}, \dots, \gamma_{Lj})^T$. Then we can apply the same group Lasso procedure for variable selection and structure identification with

$$\begin{aligned} \mathbf{W}_i(t) &= (\mathbf{B}^T(X_{i1}(t)), \dots, \mathbf{B}^T(X_{ip}(t)))^T, \quad \gamma_{-1} = (\gamma_{-11}^T, \dots, \gamma_{-1p}^T)^T, \\ P_1(\gamma_{-1}) &= \sum_{j=1}^p |\gamma_{2j}| + \sum_{j=1}^p |\gamma_{-2j}|, \\ P_h(\gamma_{-1}) &= \sum_{j=1}^p (|\gamma_{2j}|^q + |\gamma_{-2j}|^q)^{1/q} + \sum_{j=1}^p |\gamma_{-2j}|, \\ P_\infty(\gamma_{-1}) &= \max_{1 \leq j \leq p} |\gamma_{2j}| \vee |\gamma_{-2j}|, \\ Q_1(\gamma_{-1}; \lambda) &= l_p(\gamma_{-1}) + \lambda P_1(\gamma_{-1}), \quad \text{and} \quad Q_h(\gamma_{-1}; \lambda) = l_p(\gamma_{-1}) + \lambda P_h(\gamma_{-1}). \end{aligned}$$

We have the same theoretical results with just conformable changes. We should be careful in the proofs of Propositions 2 and 3 as for varying coefficient models with index variables, too. We have to deal with terms like

$$n^{-1} \sum_{i=1}^n |b_{0j}(X_{il}(t))|, \quad n^{-1} \sum_{i=1}^n |b_j(X_{il}(t))|, \quad \text{and} \quad n^{-1} \sum_{i=1}^n |b_j(X_{il}(t))b_k(X_{il}(t))|$$

as compared to

$$|b_{0j}(t)|, \quad |b_j(t)|, \quad \text{and} \quad |b_j(t)b_k(t)|$$

for time-varying coefficient models. We can use exponential inequalities for generalized U-statistics as given in Gine et al.[9] instead of Lemma 4.2 in [12] in the proof of Proposition 3.

5. Numerical studies

We carried out a small simulation study for the two models in section 4 with the P_1 penalty because time-varying coefficient models and the P_h penalty are numerically intractable at present and our computational ability is limited. We used the `grpsurv` function of the package “`grpreg`” (Breheny[1]) for R in our numerical study and all the covariates are time-independent. An extensive numerical study is a topic of future research.

First we describe the data generating process of the covariates : $\{X_{ij}\}_{j=1}^q$, $\{X_{ij}\}_{j=q+1}^p$, and Z_i are mutually independent. Then X_{ij} , $j = q + 1, \dots, p$, and Z_i follow $U(0, 1)$ independently. We define $\{X_{ij}\}_{j=1}^q$ in (35).

$$X_{ij} = F(Y_{ij}), \quad j = 1, \dots, q, \quad (35)$$

where $\{Y_{ij}\}$ is a stationary Gaussian AR(1) process with $\rho = 0.3$ and $F(y)$ is the distribution function of Y_{ij} .

Next we gives the details for our varying coefficient model with an index variable Z . We took

$$\lambda_0(t) = 0.5, \quad g_1(z) = g_2(z) = 1, \quad g_3(z) = 4z, \quad g_4(z) = 4z^2.$$

The other functions are taken to be 0. Hence we have $s_c = 4$ and $s_n = 2$. Note that X_1 and X_2 are relevant for only the constant component and that X_3 and X_4 are relevant for both the constant component and the non-constant one. All the other covariates are irrelevant. We imposed no penalty on the coefficient vector for $g_0(z)$ in this simulation study. The censoring variable C_i follows the exponential distribution with mean= 1/0.85 independently of all the other variables and the censoring rate is about 20%.

Then we describe the details for our additive model. We took

$$\begin{aligned} \lambda_0(t) &= 0.5, \quad g_1(x) = g_2(x) = 2^{1/2}(x - 1/2), \\ g_3(x) &= 2^{-1/2} \cos(2\pi x) + (x - 1/2), \quad g_4(x) = \sin(2\pi x). \end{aligned}$$

The other functions are taken to be 0. Hence we have $s_c = 4$ and $s_n = 2$ and note that X_1 and X_2 are relevant for only the linear component and that X_3 and X_4 are relevant for both the linear component and the nonlinear one. All the other covariates are irrelevant. The censoring variable C_i follows the exponential distribution with mean= 1/0.80 independently of all the other variables and the censoring rate is about 30%.

When we carried out simulations, we took $p = 400$, $q = 8$, and $L = 6$. We used the quadratic spline basis and the repetition number is 500. The results are given in Tables 1 and 2. In addition to the group Lasso, we applied a threshold method in (21) with $t_\lambda = 0.1$. In the tables, $t_\lambda = 0$ means the group Lasso and $t_\lambda = 0.1$ means this threshold group Lasso. In the tables, Failure, Correct, and False respectively stand for

Failure: The rate of relevant covariates that are not chosen wrongly,

Correct: The rate of correct decisions,

False: The rate of irrelevant covariates that are wrongly chosen.

As for the tuning parameter λ , we tried several values and found variable selection and structure identification are sensitive to this λ . We presented one of the good results for each model here. In Table 2, we sometimes missed the linear components of X_3 and X_4 . If we incorporate the assumption in (8), we will not miss these linear components. Since our procedure can be seen as a screening procedure, screening consistency or not to miss any relevant covariates is inevitable. When p is very large compared to n , it may be better to consider only variable selection based on (18) first and then apply our procedure based on some weighted $P_1(\gamma)$ as in the adaptive group Lasso.

As for tuning parameter selection rules, we don't have any results on them at present although the results of Tables 1 and 2 seem to be very promising. Some rules based on BIC, the number of selected variables, analysis of solution paths, a threshold value method, or combinations of them may be possible for screening consistency, not for selection consistency. These rules are a topic of future research since our orthonormal basis method of simultaneous variable selection and structure identification for (ultra) high-dimensional Cox models has just been proposed.

$\lambda = 0.08$	X_1 and X_2		X_3 and X_4		X_5 to $X_q(q = 8)$		X_{q+1} to $X_p(p = 400)$	
$t_\lambda = 0$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
Failure	0.000	—	0.000	0.016	—	—	—	—
Correct	1.000	0.993	1.000	0.984	0.948	0.988	0.954	0.996
False	—	0.007	—	—	0.052	0.012	0.046	0.004
$t_\lambda = 0.1$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
Failure	0.001	—	0.000	0.029	—	—	—	—
Correct	0.999	0.996	1.000	0.971	0.968	0.997	0.974	0.998
False	—	0.004	—	—	0.032	0.003	0.026	0.002

Table 1: Varying coefficient model with an index variable

$\lambda = 0.1$	X_1 and X_2		X_3 and X_4		X_5 to $X_q(q = 8)$		X_{q+1} to $X_p(p = 400)$	
$t_\lambda = 0$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
Failure	0.000	—	0.065	0.000	—	—	—	—
Correct	1.000	0.900	0.935	1.000	0.994	0.932	0.997	0.926
False	—	0.100	—	—	0.006	0.068	0.003	0.074
$t_\lambda = 0.1$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
Failure	0.000	—	0.181	0.000	—	—	—	—
Correct	1.000	0.983	0.819	1.000	1.000	0.992	1.000	0.987
False	—	0.017	—	—	0.000	0.008	0.000	0.013

Table 2: Additive model

6. Proofs

We prove Propositions 1-3 and Theorem 1.

For a vector \mathbf{a} and a matrix A , $(\mathbf{a})_i$ and $(A)_{ij}$ mean the i th element of \mathbf{a} and the (i, j) element of A , respectively. We present the proofs of technical lemmas in Appendix C.

PROOF OF PROPOSITION 1. Note that

$$\begin{aligned}
& (\hat{\gamma} - \gamma^*)^T (\dot{l}_p(\hat{\gamma}) - \dot{l}_p(\gamma^*)) \\
&= \left\{ \sum_{j \in \overline{\mathcal{S}}_c} \hat{\theta}_{1j} \frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma}) + \sum_{j \in \overline{\mathcal{S}}_c} \hat{\theta}_{-1j}^T \frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma}) \right\} \\
&+ \left\{ \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} \hat{\theta}_{1j} \frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma}) + \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} \hat{\theta}_{-1j}^T \frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma}) \right\} \\
&+ \left\{ \sum_{j \in \mathcal{S}_n} \hat{\theta}_{1j} \frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma}) + \sum_{j \in \mathcal{S}_n} \hat{\theta}_{-1j}^T \frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma}) \right\} \\
&+ \{-\hat{\theta}^T (\dot{l}_p(\gamma^*))\} = E_1 + E_2 + E_3 + E_4 \geq 0. \quad (\text{say})
\end{aligned} \tag{36}$$

The last inequality follows from the convexity of $l_p(\gamma)$ and we should recall that $\hat{\theta} = \hat{\gamma} - \gamma^*$.

We evaluate E_j , $j = 1, 2, 3, 4$.

E₁ : Notice that $\hat{\gamma}_j = \hat{\theta}_j$. Then we should evaluate

$$E_{1j} = \hat{\theta}_{1j} \frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma}) + \hat{\theta}_{-1j}^T \frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma}).$$

Recalling (19), we use the results in Appendix B.

When $\hat{\gamma}_{1j} \neq 0$ and $\hat{\gamma}_{-1j} \neq 0$, we have

$$E_{1j} = -\lambda(|\hat{\theta}_{1j}|^q + |\hat{\theta}_{-1j}|^q)^{1/q} - \lambda|\hat{\theta}_{-1j}|. \quad (37)$$

When $\hat{\gamma}_{1j} \neq 0$ and $\hat{\gamma}_{-1j} = 0$, we have

$$E_{1j} = -\lambda|\hat{\theta}_{1j}|. \quad (38)$$

When $\hat{\gamma}_{1j} = 0$ and $\hat{\gamma}_{-1j} \neq 0$, we have

$$E_{1j} = -2\lambda|\hat{\theta}_{-1j}|. \quad (39)$$

From (37)-(39), we obtain

$$E_1 \leq -\lambda \sum_{j \in \overline{\mathcal{S}}_c} (|\hat{\theta}_{1j}| + |\hat{\theta}_{-1j}|). \quad (40)$$

E₂ : Notice that $\hat{\gamma}_{-1j} = \hat{\theta}_{-1j}$ and $|\frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma})| \leq \lambda$. Then we should evaluate

$$E_{2j} = \hat{\theta}_{1j} \frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma}) + \hat{\gamma}_{-1j}^T \frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma}).$$

When $\hat{\gamma}_{1j} \neq 0$ and $\hat{\gamma}_{-1j} \neq 0$, we have

$$\begin{aligned} E_{2j} &\leq \lambda|\hat{\theta}_{1j}| - \lambda(|\hat{\gamma}_{1j}|^q + |\hat{\theta}_{-1j}|^q)^{\frac{1}{q}-1}|\hat{\theta}_{-1j}|^q - \lambda|\hat{\theta}_{-1j}| \\ &\leq \lambda(|\hat{\theta}_{1j}| - |\hat{\theta}_{-1j}|). \end{aligned} \quad (41)$$

When $\hat{\gamma}_{1j} \neq 0$ and $\hat{\gamma}_{-1j} = 0$, we have

$$E_{2j} \leq \lambda|\hat{\theta}_{1j}|. \quad (42)$$

When $\hat{\gamma}_{1j} = 0$ and $\hat{\gamma}_{-1j} \neq 0$ and when $\hat{\gamma}_{1j} = 0$ and $\hat{\gamma}_{-1j} = 0$, we have

$$E_{2j} \leq \lambda|\hat{\theta}_{1j}| - 2\lambda|\hat{\theta}_{-1j}|. \quad (43)$$

From (41)-(43), we obtain

$$E_2 \leq \lambda \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} (|\hat{\theta}_{1j}| - |\hat{\theta}_{-1j}|) \leq \lambda \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} (2|\hat{\theta}_{1j}| - |\hat{\theta}_{-1j}|). \quad (44)$$

E₃ : Notice that $|\frac{\partial l_p}{\partial \gamma_{1j}}(\hat{\gamma})| \leq \lambda$ and $|\frac{\partial l_p}{\partial \gamma_{-1j}}(\hat{\gamma})| \leq 2\lambda$. Then we have

$$E_3 \leq 2\lambda \sum_{j \in \mathcal{S}_n} (|\hat{\theta}_{1j}| + |\hat{\theta}_{-1j}|). \quad (45)$$

E₄ : We have

$$E_4 \leq P_1(\hat{\boldsymbol{\theta}})D_l = (P_1(\hat{\boldsymbol{\theta}}_S) + P_1(\hat{\boldsymbol{\theta}}_{\bar{S}}))D_l. \quad (46)$$

(40), (44), (45), and (46) yield that

$$E_1 + E_2 + E_3 + E_4 \leq (2\lambda + D_l)P_1(\hat{\boldsymbol{\theta}}_S) - (\lambda - D_l)P_1(\hat{\boldsymbol{\theta}}_{\bar{S}}).$$

The first and second inequalities follow from (36) and the above inequality. The third inequality follows from the following expression of the second one.

$$P_1(\hat{\boldsymbol{\theta}}_{\bar{S}}) \leq \frac{2\lambda + D_l}{\lambda - D_l} P_1(\hat{\boldsymbol{\theta}}_S)$$

Hence the proof of the proposition is complete.

We establish the oracle inequality.

PROOF OF THEOREM 1. First we define $D(\hat{\boldsymbol{\theta}})$ by

$$D(\boldsymbol{\theta}) = \max_{i,j} \max_{0 \leq t \leq 1} |\boldsymbol{\theta}^T \mathbf{W}_i(t) - \boldsymbol{\theta}^T \mathbf{W}_j(t)|.$$

We need two lemmas.

Lemma 1.

$$D(\boldsymbol{\theta}) \leq C_W P_1(\boldsymbol{\theta})$$

Lemma 2.

$$e^{-D(\boldsymbol{\theta})} \boldsymbol{\theta}^T \ddot{l}_p(\gamma^*) \boldsymbol{\theta} \leq (\gamma^* + \boldsymbol{\theta} - \gamma^*)^T (\dot{l}_p(\gamma^* + \boldsymbol{\theta}) - \dot{l}_p(\gamma^*)) \leq e^{D(\boldsymbol{\theta})} \boldsymbol{\theta}^T \ddot{l}_p(\gamma^*) \boldsymbol{\theta}$$

Now we begin to prove the oracle inequality. If $\hat{\boldsymbol{\theta}} = 0$, the desired inequality holds. Hence we assume $\hat{\boldsymbol{\theta}} \neq 0$ and set

$$\hat{\mathbf{b}} = \frac{\hat{\boldsymbol{\theta}}}{P_1(\hat{\boldsymbol{\theta}})}.$$

We have from Proposition 1 and the definition of $P_1(\gamma)$ that

$$\widehat{\mathbf{b}} \in \Theta\left(\frac{2+\xi}{1-\xi}\right) \quad \text{and} \quad P_1(\widehat{\mathbf{b}}) = P_1(\widehat{\mathbf{b}}_S) + P_1(\widehat{\mathbf{b}}_{\overline{S}}) = 1. \quad (47)$$

When $D_l \leq \xi\lambda$, the first inequality of Proposition 1 implies that the following inequalities hold at $x = 0$ and $x = P_1(\widehat{\boldsymbol{\theta}})$.

$$\widehat{\mathbf{b}}^T \{l(\gamma^* + x\widehat{\mathbf{b}}) - l(\gamma^*)\} \quad (48)$$

$$\begin{aligned} &\leq (2+\xi)\lambda P_1(\widehat{\mathbf{b}}_S) - (1-\xi)\lambda P_1(\widehat{\mathbf{b}}_{\overline{S}}) \\ &= 3\lambda P_1(\widehat{\mathbf{b}}_S) - \lambda(1-\xi) \leq \frac{9\lambda}{4(1-\xi)} \{P_1(\widehat{\mathbf{b}}_S)\}^2. \end{aligned} \quad (49)$$

We also used (47) here.

Note that (48) is monotone increasing and continuous in x due to the convexity of $l_p(\gamma)$ and we have (49) on $[0, P_1(\widehat{\boldsymbol{\theta}})]$. Let $x_{\mathbf{b}}$ be the maximum of x satisfying

$$\widehat{\mathbf{b}}^T \{l(\gamma^* + x\widehat{\mathbf{b}}) - l(\gamma^*)\} \leq \frac{9\lambda}{4(1-\xi)} \{P_1(\widehat{\mathbf{b}}_S)\}^2 \quad (50)$$

for any $s \in [0, x]$.

If we find an upper bound of $x_{\mathbf{b}}$, say x_0 , we have $P_1(\widehat{\boldsymbol{\theta}}) \leq x_0$. Therefore we will find an upper bound of $x_{\mathbf{b}}$ as in [12].

From Lemmas 1 and 2, we have

$$\begin{aligned} x\widehat{\mathbf{b}}^T \{l(\gamma^* + x\widehat{\mathbf{b}}) - l(\gamma^*)\} &\geq x^2 \exp\{-D(x\widehat{\mathbf{b}})\} \widehat{\mathbf{b}}^T \ddot{l}_p(\gamma^*) \widehat{\mathbf{b}} \\ &\geq x^2 \exp\{-C_W x\} \widehat{\mathbf{b}}^T \ddot{l}_p(\gamma^*) \widehat{\mathbf{b}}. \end{aligned} \quad (51)$$

The definition of κ^* and (51) imply that

$$\widehat{\mathbf{b}}^T \{l(\gamma^* + x\widehat{\mathbf{b}}) - l(\gamma^*)\} \geq x \exp\{-C_W x\} \frac{(\kappa^*)^2}{s_0} \{P_1(\widehat{\mathbf{b}}_S)\}^2. \quad (52)$$

It follows from (49) and (52) that

$$\frac{9\lambda s_0 C_W}{4(1-\xi)(\kappa^*)^2} = \tau^* \geq C_W x \exp\{-C_W x\}.$$

Consequently we have from the definition of η^* and the above inequality that

$$C_W x_{\mathbf{b}} \leq \eta^* \quad \text{and} \quad \frac{\tau^*}{\eta^*} \rightarrow 1 \text{ if } \tau^* \rightarrow 0.$$

We have found that η^*/C_W is an upper bound of $x_{\mathbf{b}}$ and that $P_1(\hat{\boldsymbol{\theta}}) \leq \eta^*/C_W$.

As for the the rest of the theorem, the result on \hat{g}_{cj} is straightforward from (20). The upper bounds on $\hat{g}_{nj}(t)$ follow from (A.1), (A.4), and the following inequalities.

$$|(\hat{\gamma}_{-1j} - \gamma_{-1j}^*)^T \mathbf{B}(t)| \leq \{\lambda_{\max}(A_{-1}A_{-1}^T)\}^{1/2} |\hat{\gamma}_{-1j} - \gamma_{-1j}^*| |\mathbf{B}_0(t)| \quad \text{and} \\ |\mathbf{B}_0(t)| \leq 1$$

Recall that the properties of our basis are collected in Appendix A.

Hence the proof of the theorem is complete.

Now we prove Proposition 2.

PROOF OF PROPOSITION 2. We implicitly carry out our evaluation on $\{\bar{Y}(1) > C_Y\}$. C_1, C_2, \dots are generic positive constants and they depend only on the assumptions.

First we deal with (32), which is represented as

$$\int_0^1 \left[\frac{S_0^{(0)}(t) \{S^{(1)}(t, \gamma^*) - S_0^{(1)}(t)\}}{S^{(0)}(t, \gamma^*) S_0^{(0)}(t)} + \frac{S_0^{(1)}(t) \{S_0^{(0)}(t) - S^{(0)}(t, \gamma^*)\}}{S^{(0)}(t, \gamma^*) S_0^{(0)}(t)} \right] d\bar{N}(t). \quad (53)$$

We can rewrite the expression in (53) as

$$(53) = (I \otimes A_0) \int_0^1 \left[\frac{S_0^{(0)}(t) \{\bar{S}^{(1)}(t, \gamma^*) - \bar{S}_0^{(1)}(t)\}}{S^{(0)}(t, \gamma^*) S_0^{(0)}(t)} \right. \\ \left. + \frac{\bar{S}_0^{(1)}(t) \{S_0^{(0)}(t) - S^{(0)}(t, \gamma^*)\}}{S^{(0)}(t, \gamma^*) S_0^{(0)}(t)} \right] d\bar{N}(t) \\ = (I \otimes A_0) \Delta l_p \quad (\text{say}), \quad (54)$$

where

$$\bar{S}^{(1)}(t, \gamma) = \frac{1}{n} \sum_{i=1}^n Y_i(t) (\mathbf{X}_i(t) \otimes \mathbf{B}_0(t)) \exp\{\mathbf{W}_i^T(t) \gamma\}, \\ \bar{S}_0^{(1)}(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) (\mathbf{X}_i(t) \otimes \mathbf{B}_0(t)) \exp\{\mathbf{X}_i(t)^T \mathbf{g}(t)\}.$$

Due to the definition of γ^* , we have uniformly in t and $l(0 \leq l < p)$,

$$|S_0^{(0)}(t) - S^{(0)}(t, \gamma^*)| \leq C_1 L^{-2}, \quad C_2 \leq S_0^{(0)}(t) \wedge S^{(0)}(t, \gamma^*), \quad S_0^{(0)}(t) \vee S^{(0)}(t, \gamma^*) \leq C_3,$$

$$|(\bar{S}_0^{(1)}(t) - \bar{S}^{(1)}(t, \gamma^*))_{lL+j}| \leq C_4 L^{-2} |b_{0j}(t)|,$$

$$|(\bar{S}_0^{(1)}(t))_{lL+j}| \vee |(\bar{S}^{(1)}(t, \gamma^*))_{lL+j}| \leq C_5 |b_{0j}(t)|.$$

Now we evaluate Δl_p . Its $(lL + j)$ th element is bounded from above by

$$C_6 L^{-2} \int_0^1 |b_{0j}(t)| d\bar{N}(t). \quad (55)$$

for some positive constant C_6 . First notice that

$$\int_0^1 |b_{0j}(t)| d\bar{N}(t) = \int_0^1 |b_{0j}(t)| d\bar{M}(t) + O(L^{-1}) \quad (56)$$

uniformly in j . Then application of an exponential inequality for martingales (Lemma 2.1 in [19]) yields

$$\mathbb{P}\left(\max_{2 \leq j \leq L} \int_0^1 |b_{0j}(t)| d\bar{M}(t) > \frac{x}{L}\right) \leq LC_7 \exp\left\{-C_8 \frac{nL^{-1}x^2}{1+x}\right\}. \quad (57)$$

We used the properties of the support of the B-spline basis in (56) and (57). Taking $x = 1$ in (57), we have established

$$|\Delta l_p|_\infty \leq \frac{C_9}{L^3} \quad (58)$$

with probability larger than $1 - LC_7 \exp\left\{-2^{-1}C_8 nL^{-1}\right\}$.

From (54), (58), and (A.3), we obtain

$$P_\infty(\dot{l}_{op} - \dot{l}_p(\gamma^*)) \leq C_{10} L^{-5/2} \quad (59)$$

with probability larger than $1 - LC_7 \exp\left\{-2^{-1}C_8 nL^{-1}\right\}$.

Finally we deal with (31) by exploiting the same exponential inequality for martingales.

For the $(lL + j)$ th element with $j = 1$, we have

$$\mathbb{P}\left(|(\dot{l}_{op})_{lL+j}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 2 \exp\left\{-\frac{C_{11}x^2 \log n}{x(n^{-1} \log n)^{1/2} + 1}\right\}. \quad (60)$$

For the $(lL + j)$ th element with $j \geq 2$, we have

$$\mathbb{P}\left(|(\dot{l}_{op})_{lL+j}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 2 \exp\left\{-\frac{C_{12}x^2 \log n}{x(n^{-1}L \log n)^{1/2} + 1}\right\}. \quad (61)$$

We used the fact that

$$\int_0^1 b_j^2(t) \lambda_0(t) dt \leq C_\lambda \mathbf{a}_{0j}^T \Omega_0 \mathbf{a}_{0j} = O(L^{-1}) \quad (62)$$

when we evaluated the predictable variation process.

It follows from (60) and (61), that

$$P_\infty(\dot{l}_{op}) \leq x(\log n)^{1/2} n^{-1/2} \quad (63)$$

with probability larger than

$$1 - 2pL \exp\left\{-\frac{C_{13}x^2 \log n}{x(n^{-1}L \log n)^{1/2} + 1}\right\}. \quad (64)$$

Hence the desired result follows from (29), (59), and (63) and the proof of the proposition is complete.

Finally we give the proof of Proposition 3.

PROOF OF PROPOSITION 3. C_1, C_2, \dots are generic positive constants and they depend only on the assumptions. We use the following lemma, which is a version of Lemma 4.1(ii) in [12].

Lemma 3.

$$\begin{aligned} \kappa^2(\zeta, \Sigma_1) &\geq \kappa^2(\zeta, \Sigma_2) - s_0(1 + \zeta)^2 L \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}| \\ RE^2(\zeta, \Sigma_1) &\geq RE^2(\zeta, \Sigma_2) - s_0(1 + \zeta)^2 L \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}| \end{aligned}$$

When $\Sigma_2 - \Sigma_1$ is *n.n.d.*, we can replace $\Sigma_2 - \Sigma_1$ in the above inequalities with Δ such that $\Delta - (\Sigma_2 - \Sigma_1)$ is *n.n.d.*

We implicitly carry out our evaluation on $\{\overline{Y}(1) > C_Y\}$. First we outline the proof and then give the details.

Define $\tilde{\Sigma}_0$ by

$$\tilde{\Sigma}_0 = \int_0^1 V_n(t, \gamma^*) S_0^{(0)}(t) \lambda_0(t) dt \quad (65)$$

and set

$$\Delta_1 = \ddot{l}_p(\gamma^*) - \tilde{\Sigma}_0 = \int_0^1 V_n(t, \gamma^*) d\overline{M}(t). \quad (66)$$

We treat Δ_1 by using the exponential inequalities for martingales.

Next define $\tilde{\Sigma}$ by

$$\tilde{\Sigma} = \int_0^1 V_n(t, \gamma^*) S^{(0)}(t, \gamma^*) \lambda_0(t) dt$$

and set $\Delta_2 = \tilde{\Sigma}_0 - \tilde{\Sigma}$. Since

$$|\mathbf{W}_i^T(t) \gamma^* - \mathbf{X}_i^T(t) \mathbf{g}(t)| \leq C_X C_{approx} L^{-2}$$

and we can use the results on predictable variation process in evaluating Δ_1 , we can easily prove

$$\max_{j,k} |(\Delta_2)_{jk}| \leq C_1 L^{-3}. \quad (67)$$

We omit the details for (67) in this paper.

Define $\hat{\Sigma}$ by

$$\hat{\Sigma} = \int_0^1 \hat{G}_Y(t) \lambda_0(t) dt, \quad (68)$$

where

$$\begin{aligned} \hat{G}_Y(t) &= \frac{1}{n} \sum_{i=1}^n Y_i(t) \{ \mathbf{W}_i(t) - \overline{\mathbf{W}}_Y(t) \}^{\otimes 2}, \\ \overline{\mathbf{W}}_Y(t) &= \frac{n^{-1} \sum_{i=1}^n Y_i(t) \mathbf{W}_i(t)}{n^{-1} \sum_{i=1}^n Y_i(t)}. \end{aligned}$$

Then by just following the arguments on pp.1161-1162 of [12] with a sufficiently small M , we obtain

$$\tilde{\Sigma} - \exp\{-C_X C_g\} \{1 + O(L^{-2})\} \hat{\Sigma} \text{ is n.n.d.} \quad (69)$$

Finally we recall the definitions of $\bar{\Sigma}$, $\bar{G}_Y(t)$, and $\boldsymbol{\mu}_Y(t)$ in Proposition 3 and set

$$\Delta_3 = \hat{\Sigma} - \bar{\Sigma} = - \int_0^1 \bar{Y}(t) \{ \bar{\mathbf{W}}_Y(t) - \boldsymbol{\mu}_Y(t) \}^{\otimes 2} \lambda_0(t) dt \quad (70)$$

and $\Delta_4 = \bar{\Sigma} - E\{\bar{\Sigma}\}$. Then we evaluate

$$\max_{j,k} |(\Delta_3)_{jk}| \quad \text{and} \quad \max_{j,k} |(\Delta_4)_{jk}|.$$

Now we give the details for Δ_1 , Δ_3 , and Δ_4 .

Δ_1 : We denote the $(jL + l, kL + m)$ element of $V_n(t, \boldsymbol{\gamma}^*)$ by $v_{jL+l, kL+m}(t)$. Then we have

$$v_{jL+l, kL+m}(t) = (S^{(2)}(t, \boldsymbol{\gamma}^*))_{jL+l, kL+m} - \frac{(S^{(1)}(t, \boldsymbol{\gamma}^*))_{jL+l} (S^{(1)}(t, \boldsymbol{\gamma}^*))_{kL+m}}{S^{(0)}(t, \boldsymbol{\gamma}^*)} \quad (71)$$

and it is easy to see that $|v_{jL+l, kL+m}(t)|$ is uniformly bounded in j, k, l, m , and t . Besides,

$$(S^{(2)}(t, \boldsymbol{\gamma}^*))_{jL+l, kL+m} \leq C_2 \begin{cases} L^{-1}, & l = m = 1 \\ L^{-1/2} |b_l(t)|, & l \geq 2, m = 1 \\ L^{-1/2} |b_m(t)|, & l = 1, m \geq 2 \\ |b_l(t)| |b_m(t)|, & l \geq 2, m \geq 2 \end{cases} \quad (72)$$

and

$$(S^{(1)}(t, \boldsymbol{\gamma}^*))_{jL+l} \leq C_3 \begin{cases} L^{-1/2}, & l = 1 \\ |b_l(t)|, & l \geq 2 \end{cases}. \quad (73)$$

By (71)-(73) and some calculation, we evaluate the predictable variation process of Δ_1 and obtain

$$\int_0^1 |v_{jL+l, kL+m}(t)|^2 d < \bar{M}, \bar{M} > (t) \leq \frac{C_4}{n} \int_0^1 |v_{jL+l, kL+m}(t)| \lambda_0(t) dt \leq \frac{C_5}{nL}, \quad (74)$$

where $< \bar{M}, \bar{M} > (t)$ is the predictable variation process of $\bar{M}(t)$. We used (62) here.

Thus we have from the exponential inequality for martingales that

$$\mathbb{P}\left(\max_{j,k} |(\Delta_1)_{jk}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 2(pL)^2 \exp\left\{-\frac{C_6 x^2 \log n}{x(\log n)^{1/2}(n^{-1}L)^{1/2} + 1}\right\}. \quad (75)$$

Δ_3 : Notice that $\bar{\Sigma} - \hat{\Sigma}$ is n.n.d. Therefore instead of Δ_3 , we treat

$$\begin{aligned} \Delta'_3 &= \frac{1}{C_Y} \int_0^1 \{\bar{Y}(t)\}^2 \{\bar{\mathbf{W}}_Y(t) - \boldsymbol{\mu}_Y(t)\}^{\otimes 2} \lambda_0(t) dt \\ &= \frac{1}{C_Y} \int_0^1 \left[n^{-1} \sum_{i=1}^n \left\{ \mathbf{W}_i(t) - Y_i(t) \boldsymbol{\mu}_Y(t) \right\} \right]^{\otimes 2} \lambda_0(t) dt. \end{aligned}$$

We evaluate $(\Delta'_3)_{kl} = (C_Y n^2)^{-1} \sum_{i,j} f_{ij}$, where $\boldsymbol{\mu}_Y(t) = (\mu_{Y1}(t), \dots, \mu_{Yp}(t))^T$ and

$$f_{ij} = \int_0^1 \{W_{ik}(t) - Y_i(t) \mu_{Yk}(t)\} \{W_{jl}(t) - Y_j(t) \mu_{Yl}(t)\} \lambda_0(t) dt.$$

Note that $|f_{ij}| \leq C_7 L^{-1}$. Thus by applying Lemma 4.2 in [12], we obtain

$$\mathbb{P}\left(\max_{k,l} |(\Delta'_3)_{kl}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 5(pL)^2 \exp\left\{-\frac{C_8 x(n \log n)^{1/2}}{x^{1/2}(n^{-1} \log n)^{1/4} + 1}\right\}. \quad (76)$$

Δ_4 : Note that

$$\begin{aligned} (\bar{\Sigma})_{kl} &= \frac{1}{n} \sum_{i=1}^n \int_0^1 Y_i(t) \{W_{ik}(t) - \mu_{Yk}(t)\} \{W_{il}(t) - \mu_{Yl}(t)\} \lambda_0(t) dt \quad \text{and} \\ \left| \int_0^1 Y_i(t) \{W_{ik}(t) - \mu_{Yk}(t)\} \{W_{il}(t) - \mu_{Yl}(t)\} \lambda_0(t) dt \right| &\leq C_9 L^{-1}. \end{aligned}$$

Applying Bernstein's inequality to $(\bar{\Sigma})_{kl}$, we have

$$\mathbb{P}\left(|(\Delta_4)_{kl}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 2 \exp\left\{-\frac{C_{10} x^2 \log n}{x(n^{-1} \log n)^{1/2} + 1}\right\}.$$

Consequently we have

$$\mathbb{P}\left(\max_{k,l} |(\Delta_4)_{kl}| \geq \frac{x(\log n)^{1/2}}{\sqrt{nL}}\right) \leq 2(pL)^2 \exp\left\{-\frac{C_{10} x^2 \log n}{x(n^{-1} \log n)^{1/2} + 1}\right\}. \quad (77)$$

By combining (66), (67), (69), (70) and (75)-(77) and exploiting Lemma 3, we obtain the desired results. Hence the proof of the proposition is complete.

7. Concluding remarks

We proposed an orthonormal basis approach for simultaneous variable selection and structure identification for varying coefficient Cox models. We have derived an oracle inequality for the group Lasso procedure and our method and theory also apply to additive Cox models. These models are among important structured nonparametric regression models. This orthonormal basis approach can be used for the adaptive group Lasso procedure. We presented some preliminary simulation results in this paper. Extensive numerical examinations and screening-consistent selection rule for λ are topics of future research.

Acknowledgments

This research is supported by JSPS KAKENHI Grant Number JP 16K05268.

Appendix A. Construction and properties of basis functions

We describe how to construct $\overline{\mathbf{B}}(t)$, the properties of $\overline{\mathbf{B}}(t)$, and the approximations to $\mathbf{g}(t)$. Set

$$\Omega_0 = \int_0^1 \mathbf{B}_0(t) \mathbf{B}_0^T(t) dt \quad \text{and} \quad \overline{\Omega} = \int_0^1 \overline{\mathbf{B}}(t) \overline{\mathbf{B}}^T(t) dt.$$

First we describe how to construct A_0 and $\overline{\mathbf{B}}(t)$. Set

$$b_1(t) = 1/\sqrt{L} \quad \text{and} \quad b_2(t) = \sqrt{12L^{-1}}(t - 1/2)$$

and define a inner product on the L_2 function space on $[0, 1]$ by

$$(g_1, g_2) = \int_0^1 g_1(t) g_2(t) dt.$$

Then we have

$$\|b_1\|^2 = \|b_2\|^2 = L^{-1} \quad \text{and} \quad (b_1, b_2) = 0.$$

Note that there is some L -dimensional vector \mathbf{a}_{02} satisfying $b_2(t) = \mathbf{a}_{02}^T \mathbf{B}_0(t)$.

We can obtain b_j , $j = 3, \dots, L$, by just applying the Gram-Schmidt orthonormalization to $(L - 2)$ elements of $\mathbf{B}_0(t)$ with the normalization of $\|b_j\|^2 = L^{-1}$. Since every $b_j(t)$ is a linear combination of $\mathbf{B}_0(t)$, we have

$$\overline{\mathbf{B}}(t) = A_0 \mathbf{B}(t).$$

Hence we have

$$\overline{\Omega} = A_0 \Omega_0 A_0^T = \begin{pmatrix} 1/L & \mathbf{0}^T \\ \mathbf{0} & \int \mathbf{B}(t) \mathbf{B}^T(t) dt \end{pmatrix} = \begin{pmatrix} 1/L & \mathbf{0}^T \\ \mathbf{0} & A_{-1} \Omega_0 A_{-1}^T \end{pmatrix} = \frac{1}{L} I. \quad (\text{A.1})$$

It is known that for some positive constants C_1 and C_2 , we have

$$\frac{C_1}{L} \leq \lambda_{\min}(\Omega_0) \leq \lambda_{\max}(\Omega_0) \leq \frac{C_2}{L} \quad (\text{A.2})$$

See Huang et al.[14] for more details.

Thus (A.1) and (A.2) imply that

$$C_3 \leq \lambda_{\min}(A_0 A_0^T) \leq \lambda_{\max}(A_0 A_0^T) \leq C_4 \quad (\text{A.3})$$

and

$$C_5 \leq \lambda_{\min}(A_{-1} A_{-1}^T) \leq \lambda_{\max}(A_{-1} A_{-1}^T) \leq C_6 \quad (\text{A.4})$$

for some positive constants C_3 , C_4 , C_5 , and C_6 . Note that (A.3) implies that

$$C_3 \leq \lambda_{\min}(A_0^T A_0) \leq \lambda_{\max}(A_0^T A_0) \leq C_4.$$

On the other hand, the definition of $\mathbf{B}_0(t)$, (A.1), and (A.4) imply that

$$\int_0^1 b_j(t) dt = 0, \text{ for } j = 2, \dots, L, \quad \text{and} \quad \sup_{2 \leq j \leq L} \|b_j\|_{\infty} = O(1). \quad (\text{A.5})$$

Besides, we have for $\gamma_j = (\gamma_{1j}, \gamma_{-1j}^T)^T \in R^L$,

$$\begin{aligned} \gamma_j^T \overline{\mathbf{B}}(t) &= \gamma_j^T A_0 \mathbf{B}_0(t) \quad \text{and} \\ |\gamma_j^T \overline{\mathbf{B}}(t)| &\leq (\gamma_j^T A_0 A_0^T \gamma_j)^{1/2} |\mathbf{B}_0(t)| \leq C_7 |\gamma_j| \end{aligned} \quad (\text{A.6})$$

uniformly on $[0, 1]$ for some positive constant C_7 . Note that we used (A.3) and the local property of $\mathbf{B}_0(t)$ to derive (A.6).

Next we consider the approximations to $\mathbf{g}(t)$. From Corollary 6.26 in [18] and Assumption G, there exist $\gamma_{0j}^* \in R^L$, $j = 1, \dots, p$, satisfying

$$\sum_{j=1}^p \|g_j - \mathbf{B}_0^T \gamma_{0j}^*\|_\infty \leq \frac{C_{approx}}{2L^2}, \quad (\text{A.7})$$

where C_{approx} depends on C_g .

In this paper, we use $\overline{\mathbf{B}}(t)$ instead of $\mathbf{B}_0(t)$. Then

$$\begin{aligned} \mathbf{B}_0^T(t) \gamma_{0j}^* &= \overline{\mathbf{B}}^T(t) (A_0^T)^{-1} \gamma_{0j}^* = \overline{\mathbf{B}}^T(t) \overline{\gamma}_j^* \quad (\text{say}) \\ &= \overline{\mathbf{B}}^T(t) \begin{pmatrix} \overline{\gamma}_{1j}^* \\ \overline{\gamma}_{-1j}^* \end{pmatrix} \quad (\text{say}). \end{aligned}$$

Noticing

$$\begin{aligned} \sum_{j=1}^p \left| \int_0^1 g_j(t) dt - \frac{\overline{\gamma}_{1j}^*}{L^{1/2}} - \int_0^1 \overline{\gamma}_{-1j}^{*T} \mathbf{B}(t) dt \right| \\ = \sum_{j=1}^p |g_{cj} - L^{-1/2} \overline{\gamma}_{1j}^*| \leq \frac{C_{approx}}{2L^2}, \end{aligned}$$

we take $\gamma_j^* = 0$ for $\overline{\mathcal{S}}_c$,

$$\begin{aligned} \gamma_{1j}^* &= L^{1/2} g_{cj} \quad \text{and} \quad \gamma_{-1j}^* = 0 \quad \text{for } j \in \mathcal{S}_c \cap \overline{\mathcal{S}}_n, \\ \gamma_{1j}^* &= L^{1/2} g_{cj} \quad \text{and} \quad \gamma_{-1j}^* = \overline{\gamma}_{-1j}^* \quad \text{for } j \in \mathcal{S}_n. \end{aligned} \quad (\text{A.8})$$

Then from (A.7), we have

$$\sum_{j=1}^p \|g_j - \overline{\mathbf{B}}^T \gamma_j^*\|_\infty \leq \frac{C_{approx}}{L^2} \quad (\text{A.9})$$

and uniformly in j ,

$$\begin{aligned} \|g_j\|^2 &= |g_{cj}|^2 + \|g_{nj}\|^2 = \gamma_j^{*T} \overline{\Omega} \gamma_j^* + O(L^{-4}) \\ &= \frac{|\gamma_{1j}^*|^2}{L} + \gamma_{-1j}^{*T} \int_0^1 \mathbf{B}(t) \mathbf{B}^T(t) dt \gamma_{-1j}^* + O(L^{-4}) \\ &= \frac{|\gamma_{1j}^*|^2}{L} + \frac{|\gamma_{-1j}^*|^2}{L} + O(L^{-4}). \end{aligned}$$

We also have

$$|g_{cj}|^2 = \frac{|\gamma_{1j}^*|^2}{L} \quad \text{and} \quad \|g_{nj}\|^2 = \frac{|\gamma_{-1j}^*|^2}{L} + O(L^{-4}). \quad (\text{A.10})$$

Appendix B. Subgradients

We give $\nabla_j P_1(\gamma)$ and $\nabla_j P_h(\gamma)$ just for reference.
For $\nabla_j P_1(\gamma)$, we have

$$\nabla_j |\gamma_{1j}| = \begin{cases} \text{sign}(\gamma_{1j}), & |\gamma_{1j}| \neq 0 \\ \epsilon_{1j}, & \gamma_{1j} = 0 \end{cases}$$

and

$$\nabla_j |\gamma_{-1j}| = \begin{cases} \gamma_{-1j}/|\gamma_{-1j}|, & |\gamma_{-1j}| \neq 0 \\ \epsilon_{-1j}, & \gamma_{-1j} = 0 \end{cases},$$

where $|\epsilon_{1j}| \leq 1$ and $|\epsilon_{-1j}| \leq 1$.

Next we deal with $\nabla_j P_h(\gamma)$. Recall that

$$\nabla_j P_h(\gamma) = \nabla_j (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} + \nabla_j |\gamma_{-1j}|.$$

Set

$$\nabla_j (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} = \begin{pmatrix} d_{1j} \\ \mathbf{d}_{-1j} \end{pmatrix},$$

where $d_{1j} \in R$ and $\mathbf{d}_{-1j} \in R^{L-1}$.

When $|\gamma_{1j}| = 0$ and $|\gamma_{-1j}| = 0$,

$$d_{1j} = \epsilon_{1j} \quad \text{and} \quad \mathbf{d}_{-1j} = \epsilon_{-1j},$$

where $|\epsilon_{1j}| \leq a$ and $|\epsilon_{-1j}| \leq b$ such that (a, b) satisfies $(1 + t^q)^{1/q} \geq a + bt$ for any $t \geq 0$. This follows from the definition of subgradient and we note that $0 \leq a \leq 1$ and $0 \leq b \leq 1$.

When $|\gamma_{1j}| \neq 0$ and $|\gamma_{-1j}| = 0$,

$$d_{1j} = \text{sign}(\gamma_{1j}) \quad \text{and} \quad \mathbf{d}_{-1j} = 0.$$

When $|\gamma_{1j}| = 0$ and $|\gamma_{-1j}| \neq 0$,

$$d_{1j} = 0 \quad \text{and} \quad \mathbf{d}_{-1j} = \gamma_{-1j}/|\gamma_{-1j}|. \tag{B.1}$$

This property is essential to hierarchical selection for g_{cj} and $g_{nj}(t)$. See [27].

When $|\gamma_{1j}| \neq 0$ and $|\gamma_{-1j}| \neq 0$,

$$d_{1j} = (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{\frac{1}{q}-1} \text{sign}(\gamma_{1j}) |\gamma_{1j}|^{q-1}$$

and

$$\mathbf{d}_{-1j} = (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{\frac{1}{q}-1} \frac{\gamma_{-1j}}{|\gamma_{-1j}|} |\gamma_{-1j}|^{q-1}.$$

Appendix C. Proofs of technical lemmas

PROOF OF LEMMA 1. From the definitions of $\overline{\mathbf{B}}(t)$ and $\mathbf{W}_i(t)$, we have

$$\boldsymbol{\theta}^T(\mathbf{W}_i(t) - \mathbf{W}_j(t)) = \boldsymbol{\theta}^T(I_p \otimes A_0)(\mathbf{X}_i(t) \otimes \mathbf{B}_0(t) - \mathbf{X}_j(t) \otimes \mathbf{B}_0(t)). \quad (\text{C.1})$$

Notice that for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_p^T)^T$,

$$|\boldsymbol{\theta}_k^T A_0 \mathbf{B}_0(t)| \leq |A_0^T \boldsymbol{\theta}_k| \leq \{\lambda_{\max}(A_0 A_0^T)\}^{1/2} |\boldsymbol{\theta}_k|. \quad (\text{C.2})$$

Here we used that $|\mathbf{B}_0(t)| \leq 1$.

Consequently (C.1) and (C.2) yield that

$$\begin{aligned} & |\boldsymbol{\theta}^T(\mathbf{W}_i(t) - \mathbf{W}_j(t))| \\ & \leq \sum_{k=1}^p |X_{ik}(t) - X_{jk}(t)| |\boldsymbol{\theta}_k^T A_0 \mathbf{B}_0(t)| \\ & \leq 2C_X \{\lambda_{\max}(A_0 A_0^T)\}^{1/2} \sum_{k=1}^p |\boldsymbol{\theta}_k| \leq C_W P_1(\boldsymbol{\theta}). \end{aligned}$$

Hence the proof is complete.

PROOF OF LEMMA 2. This lemma is just a version of Lemma 3.2 in [12]. We can verify this lemma in the same way by taking

$$a_i(t) = \boldsymbol{\theta}^T \{\mathbf{W}_i(t) - \widetilde{\mathbf{W}}_n(t, \boldsymbol{\gamma}^*)\} \quad \text{and} \quad w_i(t) = Y_i(t) \exp\{\boldsymbol{\gamma}^{*T} \mathbf{W}_i(t)\}$$

in the proof. The details are omitted. Hence the proof is complete.

PROOF OF LEMMA 3. This is almost proved in [12]. We should just note that

$$\begin{aligned} |\boldsymbol{\gamma}^T(\Sigma_1 - \Sigma_2)\boldsymbol{\gamma}| & \leq |\boldsymbol{\gamma}|_1^2 \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}| \leq L \{P_1(\boldsymbol{\gamma})\}^2 \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}|, \\ P_1(\boldsymbol{\gamma}) & \leq (1 + \zeta) P_1(\boldsymbol{\gamma}_S), \quad \text{and} \quad P_1(\boldsymbol{\gamma}_S) \leq s_0^{1/2} |\boldsymbol{\gamma}|. \end{aligned}$$

When $\Sigma_2 - \Sigma_1$ is n.n.d., we have

$$|\boldsymbol{\gamma}^T(\Sigma_1 - \Sigma_2)\boldsymbol{\gamma}| \leq \boldsymbol{\gamma}^T \Delta \boldsymbol{\gamma} \leq L \{P_1(\boldsymbol{\gamma})\}^2 \max_{j,k} |(\Delta)_{jk}|.$$

Hence the proof is complete.

Appendix D. Derivatives of the B-spline basis

In this section, we examine properties of

$$\int_0^1 \mathbf{B}'_0(t)(\mathbf{B}'_0(t))^T dt$$

and describe why we have adopted the orthogonal decomposition approach while the other authors have considered the L_2 norm of the estimated derivatives when they deal with structure identification for additive models or partially linear additive models.

We take a function $g_A(t)$ on $[0, 1]$ defined by

$$g_A(t) = \sin(2\pi At)$$

for $A \rightarrow \infty$ sufficiently slowly. Then it is easy to see

$$\|g_A\|^2 \sim 1, \quad \|g'_A\|^2 \sim A^2, \quad \text{and} \quad \|g''_A\|^2 \sim A^4.$$

On the other hand, we can approximate this $g(t)$ by $\mathbf{B}_0(t)\gamma_A$ accurately enough and we have

$$\gamma_A^T \Omega_0 \gamma_A \sim 1, \quad |\gamma_A|^2 \sim L, \quad \text{and} \quad \gamma_A^T \int_0^1 \mathbf{B}'_0(t)(\mathbf{B}'_0(t))^T dt \gamma_A \sim A^2 \rightarrow \infty.$$

This means some eigenvalues of $\int_0^1 \mathbf{B}'_0(t)(\mathbf{B}'_0(t))^T dt$ have the order larger than L^{-1} . Hence we cannot follow the proofs in the papers based on the L_2 norm of the estimated derivatives because the present eigenvalue property violates their assumptions on matrices similar to

$$\int_0^1 \mathbf{B}''_0(t)(\mathbf{B}''_0(t))^T dt.$$

The above matrix also should have some larger eigenvalues. Besides, it is more difficult to estimate the derivatives of the coefficient functions. This is why we have adopted the orthogonal decomposition approach. Zhang et al.[23] is based on the smoothing spline method and it is difficult to apply their ingenious approach to the loss function other than the L_2 loss function.

Appendix E. Proofs for other models

We outline necessary changes in the proofs for the former model in section 4 since both models in the section can be treated in almost the same way as the time-varying coefficient model. Especially, almost no change is necessary to the proofs of Proposition 1 and Theorem 1.

We assume standard assumptions for varying coefficient models here.

Proof of Proposition 2) The poof consists of (55)-(59) and (60)-(64).

(55)-(59): Note that $|b_{0j}(t)|$ is replaced with $n^{-1} \sum_{i=1}^n |b_{0j}(Z_i(t))|$. When we evaluate the predicable variation process in (57),

$$\int_0^1 |b_{0j}(t)|^2 \lambda_0(t) dt \leq C \int_0^1 |b_{0j}(t)| \lambda_0(t) dt$$

is replaced with

$$\int_0^1 \left\{ n^{-1} \sum_{i=1}^n |b_{0j}(Z_i(t))| \right\}^2 \lambda_0(t) dt \leq C \int_0^1 n^{-1} \sum_{i=1}^n b_{0j}^2(Z_i(t)) \lambda_0(t) dt. \quad (\text{E.1})$$

We can evaluate the second term in (E.1) by using Bernstein's inequality and

$$\mathbb{E} \left\{ n^{-1} \int_0^1 \sum_{i=1}^n b_{0j}^2(Z_i(t)) \lambda_0(t) dt \right\} = \int_0^1 \mathbb{E} \{ b_{0j}^2(Z_1(t)) \} \lambda_0(t) dt = O(L^{-1}).$$

(60)-(64): When we apply the martingale exponential inequality, (62) is replaced with

$$\frac{1}{n} \sum_{i=1}^n \int_0^1 b_j^2(Z_i(t)) \lambda_0(t) dt.$$

We can evaluate this expression by using Bernstein's inequality and

$$\begin{aligned} \mathbb{E} \left\{ \int_0^1 b_j^2(Z_1(t)) \lambda_0(t) dt \right\} &\leq C \mathbf{a}_{0j}^T \int_0^1 \mathbb{E} \{ \mathbf{B}_0(Z_1(t)) (\mathbf{B}_0(Z_1(t)))^T \} \lambda_0(t) dt \mathbf{a}_{0j} \\ &= O(L^{-1}). \end{aligned}$$

We need some assumptions for $\mathbb{E} \{ \mathbf{B}_0(Z_1(t)) (\mathbf{B}_0(Z_1(t)))^T \}$ as for Ω_0 in Appendix A.

Proof of Proposition 3) The proof consists of evaluating Δ_1 , Δ_3 , and Δ_4 .

Δ_1 : We should just follow the line of (60)-(64).

Δ_3 : This is almost a U-statistic and we can also apply the exponential inequality for U-statistics as (3.5) in [9] to the part of a U-statistic.

Δ_4 : This is a sum of bounded independent random variables and we can deal with this by applying Bernstein's inequality.

References

- [1] Breheny, P. The R package “grpreg” : Regularization Paths for Regression Models with Grouped Covariates. Version 3.0.1.(2016).
- [2] Bickel, P. J., Ritov, Y. A., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 37(2009) 1705-1732.
- [3] Bradic, J., Fan, J., and Jiang, J. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics* 39(2011) 3092-3120.
- [4] Bradic, J. and Song, R. Structured estimation for the nonparametric Cox model. *Electronic Journal of Statistics* 9(2015) 492-534.
- [5] Bühlmann, P. and van de Geer, S. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. 2011.
- [6] Cai, J., Fan, J., Zhou, H., and Zhou, Y. Hazard models with varying coefficients for multivariate failure time data. *The Annals of Statistics* 35(2007) 324-354.
- [7] Cai, Z. and Sun, Y. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* 30(2003) 93-111.
- [8] Cox, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 34(1972) 187-220.
- [9] Giné, E., Latała, R., and Zinn, J. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II* (pp. 13-38). Birkhäuser. 2000.

- [10] Hastie, T., Tibshirani, R., and Wainwright, M. Statistical Learning with Sparsity: the Lasso and Generalizations. CRC Press. 2015.
- [11] Honda, T. and Härdle, W. K. Variable selection in Cox regression models with varying coefficients. *Journal of Statistical Planning and Inference* 148(2014) 67-81.
- [12] Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C. H. Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics* 41(2013) 1142-1165.
- [13] Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K. Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics* 28(2000) 961-999.
- [14] Huang, J. Z., Wu, C. O., and Zhou, L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14(2004) 763-788.
- [15] Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Second Edition. Wiley. 2002.
- [16] Lian, H., Lai, P., and Liang, H. Partially linear structure selection in Cox models with varying coefficients. *Biometrics* 69(2013) 348-357.
- [17] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* 39(2011) 2164-2204.
- [18] Schumaker, L. *Spline Functions: Basic Theory*, Third Edition. Cambridge University Press. 2007
- [19] van de Geer, S. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics* 23(1995) 1779-1801.
- [20] van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer. 1996.
- [21] Yan, J. and Huang, J. Model selection for Cox models with time-varying coefficients. *Biometrics* 68(2012) 419-428.

- [22] Yang, G., Yu, Y., Li, R., and Buu, A. Feature screening in ultrahigh dimensional Cox's model. Forthcoming in *Statistica Sinica*.
- [23] Zhang, H. H., Cheng, G., and Liu, Y. Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association*. 106(2012) 1099-1112.
- [24] Zhang, H. H. and Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94(2007) 691-703.
- [25] Zhang, S., Wang, L., and Lian, H. Estimation by polynomial splines with variable selection in additive Cox models. *Statistics* 48(2014) 67-80.
- [26] Zhao, J. and Leng, C. An analysis of penalized interaction models. *Bernoulli* 22(2016) 1937-1961.
- [27] Zhao, P., Rocha, G., and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* 37(2009) 3468-3497.
- [28] Zhao, S. D. and Li, Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* 105(2012) 397-411.
- [29] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(2006) 1418-1429.